DOCUMENT RESUME

ED 452 203 TM 032 488

AUTHOR	Crislip, Marian A.; Heck, Ronald H.
TITLE	Accountability, Writing Assessment, and Equity: Testing a Multilevel Model.
PUB DATE	2001-04-00
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Accountability; Elementary Education; *Elementary School Students; Equal Education; Ethnicity; *Institutional Characteristics; Language Proficiency; Learning; *Outcomes of Education; *Performance Based Assessment; Sex Differences; Socioeconomic Status; *Student Characteristics; *Writing Evaluation; Writing Tests
IDENTIFIERS	Multilevel Analysis; Stanford Achievement Tests

ABSTRACT

The purpose of this study was to compare how learning outcomes are influenced by a number of key student composition (language background, socioeconomic status, ethnicity, and gender) and school context variables on a direct writing performance assessment. The sample was randomly selected from a population of 13,604 third graders in 175 public elementary schools in Hawaii. Students completed the Stanford Achievement Test Edition 8 (indirect writing) and the Stanford Writing Assessment Edition 1 (direct writing assessment). Students included in the sample of 3,300 had completed both the direct and indirect writing assessments. The multilevel model included student-level controls and school-level controls and the student-level variables of ethnicity, language background, socioeconomic status, and gender and the school-level variables of percent of students with limited English skills and percent of low socioeconomic status students. Results show that the set of equality variables accounted for a relatively small proportion of variance at both the student and school levels on the direct writing assessment, with only slightly more than 40% of the variance in direct writing scores attributed to student composition and school context factors. Findings suggest that direct writing assessment appeared to have measured a more diverse set of skills and to yield scores that were less affected by variables that were outside the school's control. Implications for performance assessment are discussed. (Contains 2 tables and 60 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 452 203

Marian Crislip

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Accountability, Writing Assessment, and Equity: Testing a Multilevel Model

By

Marian A. Crislip, Ph.D.
Hawaii Department of Education

Ronald H. Heck, Ph.D.
University of Hawaii - Manoa

April, 2001

TM032488

Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

BEST COPY AVAILABLE

Accountability, Writing Assessment and Equity: Testing a Multilevel Model

Beginning with Horace Mann's judgments about schools in the 1840s (Campbell, 1985; Gallegos, 1994), the process of evaluating school quality and efficiency has drawn the attention of policymakers, parents, and educational personnel. More recently, our thinking about evaluating schools has gradually evolved from a static snapshot of a school's outcomes and effectiveness in the 1980s to a more recent concern with accountability for fostering school improvement (Hallinger & Heck, 1996). At the same time, Edmonds' (1979) basic premise in the "effective schools" model was that schools that made remarkable gains in student learning had common characteristics. While differences in effectiveness could be observed, however, no clear blueprint has resulted that schools might use to become more effective.

As accountability for student learning and school improvement have increased, the measures used to assess school quality have themselves come under criticism. Traditional achievement measures, (e.g., standardized multiple-choice test scores), have been purportedly biased (gender, culture) and scores are often reported without any consideration of contextual variables that influence student outcomes. Because of the contextual influences on scores (e.g., student and school composition variables), comparisons between schools should ideally be made after adjustments for differences in their contexts. The combination, range, and variety of variables and conditions offered to ensure this, however, is necessarily complex and not standardized across the states. When used for

accountability purposes, traditional standardized tests have been found to be biased, resulted in inflated gains (e.g., Lake Wobegon effects and "saw-tooth" patterns between forms) (Linn, 2000), and were not reflective of what teachers and students actually did in the classrooms.

Performance Assessment in School Accountability

Recently, performance assessment has been proposed as a means of measuring school quality because when administered under controlled conditions they may maintain standards of validity, reliability, and fairness. Such measures may tap what students can do as a result of their education and, therefore, may conform more closely to what schools teach. Over the past few years, the use of performance-based assessments in large-scale testing has dramatically increased as an alternative to the multiple-choice format for assessing student learning and monitoring school progress. In 1998, 21 states used tests that included performance tasks. By 1999, this number had increased to 34 states (Jerald, Curran, & Boser, 1999).

Performance assessments rely on samples of students' work or judgments about their performance in completing a task that are used to evaluate their thinking skills (Wiggins, 1989). In contrast to the more narrow focus on the accumulation of facts assessed through multiple-choice tests (i.e., where the student chooses the correct answer), cognitive approaches to learning encourage the development of hands-on assessments that require students to demonstrate their acquisition of problem solving, critical thinking, and application skills that are integral to conceptual understandings of core subjects (Baxter,

Shavelson, Goldman, & Pine; 1992; Klein et al., 1997; Pearson & Valencia, 1987; Resnick & Resnick, 1992; Stecher & Mitchell, 1995; Supovitz & Brennan, 1997). Across the states currently utilizing some type of performance assessment, the tasks range from composing a sentence to completing scientific experiments and writing up the results (Jerald et al., 1999).

Performance assessments are also thought to have greater utility than multiple-choice tests for helping school personnel improve their curriculum and instructional practices (Harp, 1991; Pellegrino, 1992). Proponents suggest that they provide teachers with a means of ongoing evaluation of student progress that is more closely linked to what is actually taught. Performance assessments likely broaden teachers' curriculum responsibilities, as opposed to narrowing their responsibilities to focus on "teaching to the test" (Darling-Hammond & Goodwin, 1993; Darling-Hammond, Ancess, & Falk, 1995; Firestone, Mayrowetz, & Fairman, 1998; Garcia, 1991). Because performance assessments engage students in solving real-world problems requiring the integration of knowledge and the justification of solutions, educational reformers argue that their use can provide an impetus for changing school curriculum and classroom instructional practices, as teachers must adapt to new curriculum standards that address a wider range of student skills (Darling-Hammond, 1994; Resnick & Resnick, 1992; Smith, 1996).

Pressure to change educational practices can come from a mix of policy supports and sanctions that result from student test scores. Although the current interest in performance assessment has alleviated these concerns with the

validity of standardized tests somewhat, it has not established a credible connection with school accountability. Accounting for differences in school and home environments would make the results from performance assessments more valid. Such adjustments allow one to assess the value that the formal schooling process adds to students' lives.

There is, however, little research yet on the effects of school reforms linked to performance-based tests (Firestone et al., 1998). The correspondence of performance-based assessments to schools' intentional activities to reform their curricular and instructional practices is therefore an issue that needs careful consideration, if policymakers are to hold schools accountable for implementing new sets of curriculum standards that emphasize the development of a broader range of student skills. Because these skills may not be adequately measured on other testing formats (e.g., multiple-choice), student test scores may show no improvement, despite schools' efforts to implement the new curriculum (Mayer, 1998).

Equity in Assessment

Performance assessments may also turn out to be a more equitable testing format, if they reduce differences in scores associated with student composition (e.g., gender, ethnicity, socioeconomic status) that are normally observed on standardized tests (Darling-Hammond, 1994; Klein et al., 1997; Supovitz & Brennan, 1997). The differences in achievement for certain groups of students remain an enduring dilemma in American education. Ensuring equity in the

assessment of learning is based on the belief that all students should have access to knowledge that emphasizes conceptual understandings in the core subject areas (as opposed to the memorization of facts), the ability to use that knowledge to reason and solve problems, and the ability to communicate effectively, regardless of their socioeconomic status, ethnicity, gender, geographic location, and need for special services (Garcia & Pearson, 1994; Klein et al., 1997; Porter, 1995; Wiley & Yoon, 1995).

In the past, the inappropriate use of standardized tests in making educational decisions has posed a substantial threat to the equality of educational opportunity for significant numbers of students. Standardized tests have been challenged in the courts on the basis of their content, uses, and disproportionate impact on minority students (Garcia & Pearson, 1994; McCarthy, Cambron-McCabe, & Thomas, 1998; National Center for Fair and Open Testing, 1992). For example, testing biases have led to unreliable outcomes for minority students due to inappropriate norms (Garcia & Pearson, 1994) and, subsequently, their inappropriate placement in special assistance or remedial programs (Darling-Hammond, 1994; McCarthy et al., 1998). In contrast, however, courts have held that disproportionate impact on a particular ethnic group, in and of itself, does not violate the equal protection clause, if the quality of the test and its relationship to a legitimate educational purpose can be demonstrated (Imber & Van Geel, 1993).

While there is general agreement about the strengths and limitations of different assessment formats, there is continuing discussion over the equity,

utility, and optimum mix of the various assessments in large-scale testing. These issues are central to their construct validity in measuring intended learning tasks. Whether performance-based tests will be even more widely used will depend heavily upon the quality of the tests, their demonstrated relationship to curricular and instructional goals, and costs associated with their development, implementation, and scoring (Stecher & Klein, 1997). It is therefore important to ensure that performance assessments developed for large-scale testing demonstrate valid and reliable measurement of designated learning tasks that correspond to the school's curricular standards and are fair and nondiscriminatory to the fullest extent practical and possible (Darling-Hammond, 1994; Imber & Van Geel, 1993; LaMorte, 1996; Linn, 1994; McCarthy et al., 1998; Moss, 1994; Wenglinsky, 1998). In other words, measures should be chosen that are sensitive to what school personnel are trying to teach and minimally affected by those factors that they cannot control.

Writing Performance Assessment

Writing performance assessment is relatively new in state testing, and therefore, its equity and utility for attaining educational and accountability purposes have not yet been fully demonstrated (Darling-Hammond et al., 1995; Gronlund, 1988; Linn, 1994; Manzo, 2000; Supovitz & Brennan, 1997). The purpose of our study is to compare how learning outcomes are influenced by a number of key student composition (e.g., language background, socioeconomic status, ethnicity, gender) and school context variables on a direct writing performance assessment.

A recent issue of *Education Week* (January 11, 1999) was devoted to a state-by-state description and comparison of current assessment and school accountability practices. These practices vary considerably across the 50 states (Jerald et al., 1999). A summary outlining the standardized tests and performance-based tests used in assessing K-12 student learning in the four core subject areas (language arts, math, science, social studies) indicates that writing assessment is the most common type of performance-based measure used (Jerald & Boser, 1999). Most often, writing assessment is used in addition to standardized tests, or a combination of standardized tests and performance assessments. In a few states, however, performance-based measures are now used exclusively.

Only a few states test students in grades K-2 (Jerald & Boser, 1999). Most states begin testing students in the middle elementary grades. To examine the practices more closely, 28 states currently assess third grade students' language arts skills. Sixteen of the states rely exclusively on standardized tests, 11 states provide some types of performance tasks (e.g., constructing a writing response), and one state uses both multiple-choice and performance-based measures. Twenty-seven states use at least some performance tasks in language arts at the fourth grade level, and of these, 23 states include writing assessments. For secondary grades, this general pattern is repeated, with the greatest number of states testing language arts skills in grades 8 and 10.

In the past, student writing was assessed indirectly (e.g., knowledge of writing mechanics, sentence structure, syntax, and grammar). More recently, test

developers have also included the "holistic" aspects (e.g., general merit, organization of ideas) of the writing process (Spandel & Stiggins, 1990). There appears to be a substantive difference in the validity (i.e., particularly *prima facie* validity) of direct assessments of student writing compared with those indirect assessments that use a multiple-choice format to assess students' writing skills (Greenberg, 1986; Linn, 1991; Meredith, 1984; Smith, 1978; White, 1994). A number of studies have explored the relationship between direct and indirect writing measures (Bennett, Rock, & Wang, 1991; Breland & Gaynor, 1979; Culpepper & Ramsdell, 1982; Finch, 1991; Godshalk, Swineford, & Coffman, 1966; Greenberg, Wierner, H.S., & Donovan, 1986; Haladyna, 1998; Hennings, 1996; Linn, Baker, & Dunbar, 1991; Lombard, 1988; Meredith & Williams, 1984; Smith, 1982; Spandel & Stiggins, 1990; Stiggins, 1981; White, 1994). Some researchers view constructed writing responses as complementary with the results of multiple-choice assessments (e.g., Bennett et al., 1991; Breland & Gaynor, 1979; Culpepper & Ramsdell, 1982), while the argument against a comparison of direct and indirect assessments of writing is primarily that they measure different constructs (Resnick & Resnick, 1992). Others have found some correspondence between formats, but suggested that strength of the relationship depended on the ability levels of the students (Hennings, 1996; Stephenson & Giacoboni, 1988). Few investigations of writing performance assessment exist, however, in studies of school effects and school improvement.

Because research on writing performance assessment of writing is relatively new, there is a need to identify student and school variables that might

influence the quality of students' writing. Very little investigation has been conducted to explore the relationships that student background and school contextual factors have on writing assessment. If writing is to be used as an accountability measure for determining school quality and school improvement, then it is important to know how schools' scores may be affected by the composition of their students and school-level variables. Examining how groups of students perform may also yield important information about the equity of these tests.

The utility of writing performance assessments would be enhanced if they could be shown to be less sensitive than typical standardized tests to variables that schools cannot control, while being more sensitive to their curricular and instructional processes. In the analysis that follows, we examine the relative equity associated with direct writing assessment on two levels. Equity in terms of individual comparison should focus on the performance outcomes achieved by various groups of students within schools (e.g., females, low-SES students). At the student level, we examine the extent to which individual students' backgrounds affect their performance on the writing performance assessment.

Equity in terms of school comparison should focus on how differences between schools in student composition and other school context variables may influence the school outcomes produced. At the school level, we examine the extent to which student composition, school context, and school process affect school outcomes on the performance writing assessment. Through this

examination, we may gain further understanding about the validity of using different testing formats for various educational purposes.

Method

Sample

The sample of students in this study was randomly selected from a population of 13,064 third grade students in 175 public elementary schools in Hawai'i. Third grade students were administered the Stanford Achievement Test (Gardner et al., 1985) Edition 8 and the Stanford Writing Assessment Edition 1 (Stanford Achievement Test, 1983) in the spring of 1993. Students needed to have completed both the direct writing and indirect writing assessments to be included in the sample. After eliminating a number of schools with insufficient data (e.g., new schools, K-2 school configurations, small school sizes), twenty third-grade students were randomly selected within each of the 165 remaining schools to participate in the study (N=3,300).

Other school data were compiled from the School Status and Improvement Report, used by Hawai'i Department of Education as part of the state's assessment and accountability program.

Variables in the Model

Outcomes

Direct writing. The Stanford Writing Assessment is a constructed-response assessment of student writing ability, consisting of a single draft, prompt-directed writing sample that is written within a 25 minute time period. The prompt was as follows:

A principal wants to make the school day longer. School would start each day at 8:00 a.m. and end at 4:30 p.m. Do you think this is a good idea? Why or why not?

The writing samples were analytically (trait) scored by trained readers (teams of classroom teachers) using a seven-point scale for each of the five domains (i.e., general merit, ideas, organization, words, and syntax). The scores in each domain consisted of mean scores across two readers. To monitor rater "drift," selected papers were also scored by a third member of each group. Students whose papers were marked as "off topic" were given a score and included in the data. Students whose papers were identified as "illegible" were omitted from the data set. Inter-rater reliability for the training session was .91 over a set of ten papers (Hawaii Department of Education, 1993). This reliability compared favorably with Supovitz and Brennan's (1997) study of portfolio assessment (with reported inter-rater reliability coefficients of .73 for first grade portfolios and .78 for second grade portfolios).

We used principal components analysis to create a total direct writing score for each student (i.e., mean = 50 and standard deviation = 10). The five domains comprising the writing component all had factor loadings greater than .96, and the component accounted for 94.5 percent of the total variance among the five domains.

Student-Level Variables

Student controls. Students' age (in years and partial years) and whether or not they received special education (SPED) services were included as background controls.

Equity variables. Our primary interest lies in examining the impact of student composition variables on the two sets of assessment scores. The variables were ethnicity, socioeconomic status, gender, home language, and language-related services. Student ethnicity was disaggregated into a set of dummy-coded variables with Caucasians serving as the reference group. Because no direct measure of the student's socioeconomic status was available, we used student participation in the state's federally subsidized lunch program to serve as an indirect indicator of low student socioeconomic status (low SES). The 12 home language categories included in the study were English, Cantonese, Mandarin, Ilokano, Tagalog, Cebu/Visayan, Hawaiian, Japanese, Korean, Samoan, Vietnamese, and others. Ninety percent of sample students spoke English as the first language in the home. Languages other than English were recoded into non-English speaking. A smaller percentage of this latter group (i.e., three percent of the entire sample) received support services for limited English proficiency (SLEP).

School-Level Variables

School controls. We included several school-level controls that were used in previous research on student writing (e.g., Supovitz & Brennan, 1997). These

variables were school size, staff stability (i.e., the percentage of teachers who have been at the school for five or more years), student stability (i.e., the percentage of students enrolled in the school for the entire school year), and the percentage of special education students in the school.

School equity variables. The school equity variables were the percentage of students in the school who received free/reduced lunch subsidies and the percentage of students who received services for limited English proficiency.

Writing/Language arts school improvement. We also examined the impact of schools' planned improvement efforts on student scores. Each school year, Hawai'i school administrators submit the School Status and Improvement Report (SSIR) to the state superintendent. This form is divided into (1) school context indicators, (2) school improvement process, and (3) outcomes. Context indicators present information about the school's students, instructional staff, and facilities. In the section on school improvement, each principal ranks the school's top three chosen improvement priorities and provides a brief description of the school activities (e.g., staff development, purchase of materials to support student learning) to address them. On the SSIR, 10 possible improvement priorities can be identified: student achievement, student behavior, student attitudes, school curriculum, staff development, campus facilities and appearance, parent involvement/community relations, school/community-based management, school support services, and innovative programs.

In addition to identifying the priorities for improvement, a portion of report allows the school principal to write a narrative for further clarification of the

school's plan. These reports were content analyzed by reviewing each school's narrative section for the previous two school years to determine whether or not it had identified student writing skills to be an area of improvement. For example, a school might choose to emphasize student achievement in the general section of the SSIR form and indicate in the narrative section that it "decided to focus on the improvement of our students' writing skills by incorporating writing activities across curriculum content areas" (principal comments). After completing our content analysis, we coded this variable "2" if writing had been selected as a improvement emphasis over two consecutive years, "1" if writing had been indicated as an emphasis during one of the two years, and "0" if writing had not been selected as an improvement emphasis.

Model and Analyses

In beginning to develop a performance assessment model, the complexity of educational systems across multiple levels becomes an issue—that is, children in classes, classes in schools, schools in districts. Variables at each level have the potential to impact test scores. While several problems surrounding testing practices (e.g., community factors, school characteristics) have been discussed, schools have not, as of this date, undergone an extensive and comprehensive multilevel comparative study that uses writing as a mode of performance assessment. Multilevel modeling of variables that affect writing should help policy makers determine the extent to which performance assessments are equitable across groups of students, as well as what types of school variables affect writing achievement.

The proposed multilevel conceptual model is presented in Figure 1. This model reflects several sets of variables identified in previous research as impacting on student achievement (e.g., Creemers, 1994). A series of models was proposed. We begin with a set of student-level controls (i.e., age, participation in special education) and school-level controls (e.g., school size, student and staff stability). To that preliminary model, we add the set of student-level (i.e., ethnicity, language background, socioeconomic status, gender) and school-level (i.e., percent of students with limited English skills and percent of low socioeconomic students) equity predictors. Finally, we add the school process variable (i.e., relative focus on improving school writing scores).

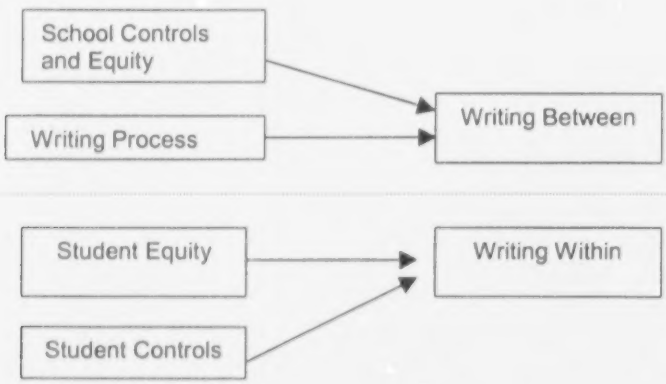

Figure 1

Proposed Multilevel Model of Variables Affecting Writing Achievement

Results

Descriptive data about the students and schools in the analysis are summarized in Table 1.

Insert Table 1 About Here

Comparing Models

One goal of our analysis was to determine how variance in writing scores would be attributed to the various controls, equity predictors, and process variables accounted for variance in students' writing scores. We investigated a series of models. The first (null) model contained no predictors, but we can use the variance components to calculate the amount of variance that is accounted for by the sets of predictors included in each successive model. For the writing model, the intraclass correlation (i.e., the proportion of variance in the outcome that lies between schools) was .24, suggesting there was considerable variability in student writing performance that might be explained by differences between schools (e.g., school context, school processes). For example, intraclass correlations on standardized tests have been reported to range from between .10 and .20 (Hill & Rowe, 1996).

The baseline model added the set of student (i.e., age, special education status) and school controls (i.e., school size, staff stability, student stability, percentage of special education students). At the individual student level, this model accounted for 5 percent of the variance in the students' writing scores, and at the school level, the baseline model accounted for 23 percent of the variance in the scores.

Next, we added the set of equity variables to the set of variables comprising the control (or baseline) model. Together, these sets of variables represent factors that are outside of the school's control. This model accounted

for 13 percent of the explained variance in writing scores at the student level and 43 percent of the explained variance in writing scores at the school level. Finally, the complete multilevel model, which added the writing improvement process variable, accounted for about 45 percent of the between-school variance in the direct writing scores (while the student-level variance remained the same at 13%).

Examining Individual Model Parameters

Next, we examined the effects of individual student- and school-level variables on the writing scores. The parameter estimates for the complete multilevel model are presented in Table 2. Because students' writing scores were standardized to have a variance equal to 100 (i.e., standard deviation equal to 10), the coefficients are readily interpretable in terms of standard deviations. For example, a coefficient of 5.00 on a dummy variable would represent a one-half standard deviation achievement difference between the two groups.

Student Variables

For the student-level controls, the gap in performance was rather strong for SPED students on the direct writing assessment (-8.204, $p < .05$), as compared to their regular education counterparts. Age, however, was not a significant predictor of performance on either assessment. For the student-level equity variables, the effect of home language itself was insignificant on the writing assessment. More specifically, however, the relatively small group of students who received language support services (SLEP) scored much lower than their non-SLEP counterparts (-8.964, $p < .05$). Low SES students scored

significantly lower on the direct writing assessment (-1.900, p< .05). Girls had significantly higher predicted writing scores than boys on both the direct assessment (2.318, p < .05). We observed only a few significant differences in performance across ethnic groups. More specifically, only Hawaiian (-2.249), Portuguese (-3.719), and Samoan (-2.054) students scored significantly lower (p < .05) than the reference group, and no groups scored significantly higher.

Insert Table 2 About Here

School Variables

We also examined the impact of the school variables on the student writing scores. For the school-level context controls and equity variables, percent of special education students (-.289), percent of SLEP students (-.137), and percent of low SES students (-.045) were negatively related to writing outcomes (p < .05). In contrast, student stability was positively related to direct writing outcomes (.233, p < .05). This latter finding suggests that schools having higher percentages of students enrolled throughout the entire school year had higher writing outcomes. Finally, the school's stated commitment to planned improvement in writing was significantly related to the direct writing scores (1.051, p < .05). More specifically, each year that a school identified it had worked on writing as a central part of its school improvement process resulted in a .11 standard deviation advantage in writing scores. This finding provided preliminary evidence that the school's planned curricular improvement in writing/language arts was related to its subsequent higher writing outcomes.

Discussion and Implications

The purposes of our study were (1) to examine the extent to which direct assessment of student writing is affected by student composition and (2) to examine their validity in comparing schools for several educational purposes (e.g., accountability, impact of curricular improvement processes). The study raises a number of policy issues concerning the relative equity and utility of this type of assessment for comparing students and schools.

To What Extent is Direct Writing Assessment More Equitable?

Proponents of performance assessments have argued that it is important to ensure that assessments developed for large-scale testing demonstrate valid and reliable measurement of designated learning tasks that correspond to the school's curricular standards and are fair and nondiscriminatory to the fullest extent practical and possible. If direct writing assessments provide a more equitable testing format, after taking into account students' age, ability differences, and school context controls, the inclusion of a set of important equity variables (i.e., gender, ethnicity, socioeconomic status, language background) should contribute little to the explanation of student performance.

Our results provide some evidence that this is indeed the case; that is, the set of equity variables accounted for a relatively small proportion of variance at both the student and school levels on the direct writing assessment. Only slightly more than 40 percent of the variance in direct writing scores was attributed to student composition and school context factors. We found that four of the equity predictors significantly impacted students' direct writing scores. These variables

were SLEP participation, low socioeconomic status, ethnicity, and gender. Similar to other research, performance assessment in writing did not reduce these differences entirely (e.g., Jovanovic et al., 1994; Peng et al., 1995; Supovitz & Brennan, 1997). These same variables also significantly impacted students' indirect writing scores. Students receiving language support services and students in special education had difficulty with the direct writing format. We suspect their low performance on the direct writing assessment is related to the difficulties of teaching writing skills to these groups of children.

The good news in our study was that the direct writing format produced fairly small achievement differences for low socioeconomic students. We also found that the direct writing assessment tended to reduce the size of achievement differences for students of some ethnic backgrounds (i.e., Hawaiian, Filipino, Samoan). These students have been previously identified as scoring below state averages over time on standardized tests (Kamehameha Schools/Bishop Estate, 1993). It is likely that the reduced gaps in achievement were due to a variety of factors including the construction of the tests themselves (Klein et al., 1997; Resnick & Resnick, 1992; Supovitz & Brennan, 1997) and other student abilities and background variables. Some of these variables may include reading ability, home environment, and the quality of the schools students attend (e.g., Lee & Bryk, 1989; Peng et al., 1995).

Are Writing Assessments Useful in Monitoring School Improvement?

Thus far, we have argued that direct writing assessment is an important means for monitoring student learning and comparing schools. We provided

evidence indicating the direct writing assessment likely measured a more diverse set of skills and students' scores were less affected by variables outside the school's control. Performance assessments have also been thought to provide a closer linkage between the learning skills emphasized in the school's curriculum and the actual assessment of those skills. While reforms over the past decade have focused on improving the school's instructional practices and developing appropriate measures for measuring student learning (Louis, Toole, & Hargreaves, 1999), the effects of school improvement efforts have been difficult to demonstrate, partly because of the greater impact of student composition and school context on student performance on standardized tests and the mismatch between what is taught and what is measured (Mayer, 1998; Wiley & Yoon, 1995).

Our results suggested that the direct writing assessment corresponded with the school's deliberate focus and energy directed toward curricular and instructional improvement, that is, schools that had concentrated on improving student writing over time had higher writing scores. This provides preliminary evidence of the writing performance assessment's construct validity in estimating the effect of schools' efforts to reform their educational practices, where such evidence has been previously lacking (Firestone et al., 1998). This effect was consistent when the writing process variable was added to the statistical model in different places; however, we found that it had to be included in the model after school SES was controlled, which suggests the greater difficulties associated with improving schools in low-SES communities.

Although our proxy measure of the school improvement process admittedly needs more refinement (as it depended on the content analysis of school narratives), it is actually surprising that this variable picks up as much variance in school means as it does. While we suggest caution in interpreting this result, from a practical perspective, the finding is encouraging because it provides evidence supporting the view that schools can undergo a specific, purposeful improvement process that helps them become more effective (Ouston, 1999). This correspondence also supports the view that writing performance assessment likely measures what students learn in the school as opposed to what inequities they bring from the home (Peng et al., 1995).

We were unable to discover what the school personnel actually did over the course of time to strengthen their writing curriculum. Unfortunately, while the school data once existed, the specific activities of schools during this time can no longer be found recorded within the Hawai'i Department of Education. It may be that a focus on writing in a school introduces a completely new element in the curriculum. This may result in different resource allocations, staff development activities, and increased opportunities for students to write. Another possibility is a district or state policy initiative, like the introduction of a new test, may result in changes to the curriculum. We deem this less likely in this case, however, because the state had administered a writing test in addition to multiple-choice tests in reading, language, and math since the early 1980s, but the results on the writing assessment were never made public.

Concluding Thoughts

Increasingly, states are adopting performance-based writing measures as part of their large-scale student testing and school accountability programs. Our initial investigation of the direct writing formats supports the continued refinement and use of performance-based writing assessments for school accountability purposes. Further research should be directed at determining how student ability and background may interact with the content and format of the writing tests developed, the correspondence and divergence in cognitive domains that each format assesses, and the optimum number of content tasks in writing that might be needed to generalize about the overall performance (i.e., most writing assessments are currently limited, consisting of one short writing prompt). The increased economic costs of developing, administering, and scoring writing performance assessments must be balanced against the social and political costs of continued over-reliance on standardized, multiple-choice tests that may be less useful in monitoring what students can do as a result of their educations and may also be biased in known and unknown ways for some groups of students. Coupled with evidence that direct writing assessment reduced achievement differences commonly observed on multiple-choice tests for some groups of students, our results also provide support for the view that writing performance assessment should be utiiized in large-scale testing because it provides a more valid comparison between schools.

Table 1
Descriptive Statistics

	Mean*		Minimum	Maximum
School Level (N=165)				
School Size	633.99	(280.51)	170.00	1561.00
%Low SES	38.49	(20.93)	2.18	90.24
%SPED	5.80	(2.95)	0.00	17.15
%SLEP	5.23	(5.24)	0.00	27.00
%Students Enrolled	91.16	(4.79)	68.59	99.81
%Staff Stability	59.00	(18.00)	0.00	100.00
Writing Plan	0.41	(0.68)	0.00	2.00
Student Level (N=3300)				
Direct Writing	50.00	(10.00)	18.18	87.08
SLEP	0.03		0.00	1.00
SPED	0.05		0.00	1.00
Age	8.89		7.48	10.55
Female	0.50		0.00	1.00
Home Language English	0.90		0.00	1.00
Native American	0.00		0.00	1.00
African American	0.02		0.00	1.00
Chinese	0.03		0.00	1.00
Filipino	0.15		0.00	1.00
Hawaiian	0.25		0.00	1.00
Hispanic	0.02		0.00	1.00
Indonesian	0.01		0.00	1.00
Japanese	0.13		0.00	1.00
Korean	0.02		0.00	1.00
Portuguese	0.02		0.00	1.00
Samoan	0.03		0.00	1.00
Other Ethnicity	0.10		0.00	1.00

Note: *Standard deviations of continuous variables are included in parentheses.

Table 2
HLM Coefficients for Full Model Explaining Third Grade Student Performance in Writing

	Writing Performance	
	Coefficient	SE
School-level Variables		
Controls		
School Size	0.000	0.001
Sped%	-0.289*	0.103
Student Stability	0.233*	0.069
Staff Stability	2.073	1.961
Equity Predictors		
Slep%	-0.137**	0.071
Low Ses%	-0.045*	0.020
School Process		
Writing Plan	1.051*	0.496
Student-level Variables		
Controls		
Age	-0.002	0.352
Sped	-8.204*	0.740
Equity Predictors		
Slep	-8.964*	1.068
Home Language English	0.168	0.663
Low SES	-1.900*	0.325
Female	2.318*	0.259
African American	-1.471	1.028
Chinese	1.114	1.149
Filipino	-0.851	0.572
Hawaiian	-2.249*	0.512
Hispanic	-1.442	1.278
Indonesian	0.411	1.363
Japanese	0.613	0.553
Korean	1.239	1.195
Native American	-0.738	2.685
Portuguese	-3.719*	1.036
Samoan	-2.054*	2.830**
Other Ethnicity	-1.005**	0.513

Notes: *p < .05, **p < .10
Writing reliability estimate = 0.86

References

Baxter, G, Shavelson, R., Goldman, S., & Pine, J. (1992). Evaluation of procedure based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1 17.

Bennett, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28(1), 77-92.

Breland, H.M., & Gaynor, J.L. (1997). A comparison of direct and indirect assessment of writing skill. Journal of Educational Measurement, 16(2), 119-128.

Campbell, R.F., Cunningham, L.L., Nystrand, R.O., & Usdan, M.D. (1985).The organization and control of American schools. Columbus, OH: Merrill.

Creemers, B.P.M. (1994). The international school effectiveness research programme ISERP, First results of the quantitative study. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Culpepper, M., & Ramsdell, R. (1982). A comparison of multiple-choice and an essay test of writing skills. Research in the Teaching of English, 16,295-297.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. Harvard Educational Review, 64(1), 5-29.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). Authentic assessment in action. New York: Teachers College Press.

Darling-Hammond, L. & Goodwin. L. (1993). Progress toward professionalism in teaching. In G. Kawelti (Ed.), Challenges and achievements of American education. Alexandria, VA: Association for Supervision and Curriculum Development, 19-52.

Edmonds, R. (1979). Effective schools for the urban poor. Educational Leadership, 37, 15-24.

Finch, F.L. (Ed.) (1991). Educational performance assessment. Chicago, IL: Riverside.

Firestone, W., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in

Maine and Maryland. Educational Evaluation and Policy Analysis, 30(2), 95-113.

Garcia, G. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. Reading Research Quarterly, 26, 371-392.

Garcia, G. & Pearson, P. (19940. Assessment and diversity. In L. Darling-Hammond (Ed.), Review of research in education, 20. Washington, DC: American Educational Research Association, 337-383.

Gardner, E.F., Rudman, H.C., Carlsen, B., & Merwin, J.C. (1998). The Stanford achievement test (8th Ed.) San Antonio, TX: The Psychological Corporation.

Gallegos, A. (1994). Meta-evaluation of school evaluation models. Studies in Education Evaluation, 20, 41-54.

Godshalk, F.I., Swineford, F., & Coffman, W.E. (1966). The measurement of writing ability. New York, NY: College Entrance Examination Board.

Greenberg, K.L., Wierner, H.S., & Donovan, R. (1986). Writing assessment: Issues and strategies. New York: Longman.

Gronland, N. (1998). Assessment of student achievement. (6th Ed.). Boston, MA: Allyn & Bacon.

Haladyna, T.M. (1998). Fidelity and proximity to criterion: When should we use multiple-choice? Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Hallinger, P., & Heck, R.H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. Educational Administration Quarterly, 32 (1), 5-44.

Harp, W. (1991). Principles of assessment and evaluation in whole language classrooms. In W. Harp (Ed.), Assessment and evaluation in whole language programs. Norwood, MA: Christopher Gordon, 35-50.

Hennings, S.S. (1996). A comparison of equating methods applied to performance based assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Hill, P.W., & Rowe, K.J. (1996). Multilevel modeling in school effectiveness research. School Effectiveness and School Improvement, 7 (1), 1-34.

Imber, M. & Van Geel, T. (1993). Education Law. New York: McGraw-Hill.

Jerald, C. & Boser, U. (January 11, 1999). Taking stock. Education Week, 18(17), 81-99.

Jerald, C., Curran, B., & Boser, U. (January 11, 1999. State of the states. Education Week, 18(17). 106.

Klein, S., Jovanovic, J., Stecher, B., McCaffrey, D., Shavelson, R., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. Educational Evaluation and Policy Analysis, 19(2), 83-97.

LaMorte, M. (1996). School law: Cases and concepts. Boston,MA: Allyn & Bacon.

Lee, V. & Bryk, A. (1989). A multilevel model of the social distribution of high school achievement. Sociology of Education, 62, 172-192.

Linn, R.L. (1994). Performance assessment: Policy, promises, and technical measurement standards. Educational Researcher, 23(9), 4-14.

Linn, R.L. (2000). Assessments and accountability. Educational Researcher, 29 (2), 4-16.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 15-21.

Lombard, J.V. (1988). An empirical comparison of a direct and an indirect method of assessing writing proficiency. Eric Document Reproduction Service ED 303 519.

Louis, K.S., Tooke, J., & Hargreaves, A. (1999). Rethinking school improvement. In J. Murphy & K. Seashore-Louis (Eds.), The handbook of research on educational administration (2nd edition.). San Francisco, CA: Jossey-Bass, 251-276.

Manzo, K. (2000). NAEP drops long-term writing data. Education Week, 29(27), 1,17.

Mayer, D.P. (1998). Do new teaching standards undermine performance on old tests? Educational Evaluation and Policy Analysis, 20(2), 53-73.

McCarthy, M., Cambron-McCabe, N., & Thomas, S. (1998). Public school law: Teachers' and students' rights (4th edition). Boston, MA: Allyn & Bacon.

Moss, P.A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

Meredith, V.H., & Williams, P.L. (1984). Issues in direct writing assessment problem identification and control. Educational Measurement, 3, 11-15.

National Center for Fair and Open Testing. (1992). K-12 testing fact sheet. Cambridge, MA: Author.

Ouston, J. (1999). School effectiveness and school improvement: Critique of a movement. In T.Bush, L. Bell, R. Bolam, R. Glatter, & P. Ribbins (Eds.), Educational management: Redefining theory, policy and practice. London: Paul Chapman Publishing, 166-177.

Pearson, P. & Valencia, S. (1987). Assessment, accountability, and professional perogative. In J. Readence & R. Baldwin (Eds.), Research in literacy: Merging perspectives. Thirty-sixth yearbook of the National Reading Conference. Rochester, NY: National Reading Conference, 3-16.

Pellegrino, J. 91992). Commentary: Understanding what we measure and measuring what we understand. In B. Gifford & M. O'Conner (Eds.), Changing assessment: Alternative views of aptitude, achievement, and instruction. Boston, MA: Kluwer, 275-294.

Porter, A. (1995). The uses and misuses of opportunity to learn standards. Educational Researcher, 24(1), 21-27.

Resnick, L. & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Conner (Eds.), Changing assessments: Alternative views of aptitude, achievement, and instruction. Boston, MA: Kluwer, 37-75.

Sammons, P., Hillman, J., & Mortimore, P. (1995). Key characteristics of effective schools: A review of school effectiveness research. London: International School Effectiveness & Improvement Centre, University of London.

Spandel, V., & Stiggins, R.J. (1990). Creating writers: Linking assessment and writing instruction. New York, London: Longman.

Smith, L.S. (1982). Investigation of writing assessment strategies: Studies in measurement and methodology: ERIC Document Reproduction Service ED 213 727.

Smith, M.L. (1996). Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program. Tempe,AZ: Southwest Educational Policy Studies, Arizona State University.

Stephenson, R.S., & Giacoboni, K.N. (1988). A comparison of 1987 results of SSAT-I writing and production writing assessment.

Stecher, B. & Klein, S. (1997). The cost of science performance assessments in large-scale testing programs. Educational Evaluation and Policy Analysis, 19(1), 1-14.

Stecher, B. & Mitchell, K. (1995). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving. (CSE Technical Report NO 400). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles.

Stiggins, R.J. (1981). A comparison of direct and indirect writing assessment methods. ERIC document Reproduction Service, ED 204 413.

Supovitz, J.A., & Brennan, R.T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. Harvard Educational Review, 67(3), 472-506.

White, E.M. (1994). Teaching and assessing writing: Recent advances in understanding, evaluation, and improving student performance. San Francisco, CA: Jossey Bass.

Wenglinsky, H. (1998). Finance equalization and within-school equity: The relationship between education spending and the social distribution of achievement. Education Evaluation and Policy Analysis, 20(4), 269-283.

Wiggins, G. (1989). Teaching to the (authentic) test. Educational Leadership, 46(4), 41-47.

Wiley, D.E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). Educational Evaluation & Policy Analysis, 17 (3), 355-70.